EP0933712

Publication Title:

Method and system for generating document summaries with navigation information

Abstract:

Abstract of EP0933712

A method and a system for generating a document summary. The method and the system extract text along with corresponding position information from a document, provide the extracted text to a document summarizer (62) and generates a summary with indicators that indicate the corresponding locations in the document from which the summary portions were extracted.

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 933 712 A2

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
04.08.1999 Bulletin 1999/31

(51) Int Cl.^{6}: **G06F 17/27**, G06F 17/30

(21) Application number: 99300218.7

(22) Date of filing: 14.01.1999

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: 30.01.1998 US 16224

(71) Applicant: XEROX CORPORATION
Rochester, New York 14644 (US)

(72) Inventor: Kupiec, Julian M.
Cupertino, California 95014 (US)

(74) Representative: Skone James, Robert Edmund
GILL JENNINGS & EVERY
Broadgate House
7 Eldon Street
London EC2M 7LH (GB)

(54) Method and system for generating document summaries with navigation information

(57) A method and a system for generating a document summary. The method and the system extract text along with corresponding position information from a document, provide the extracted text to a document summarizer (62) and generates a summary with indicators that indicate the corresponding locations in the document from which the summary portions were extracted.

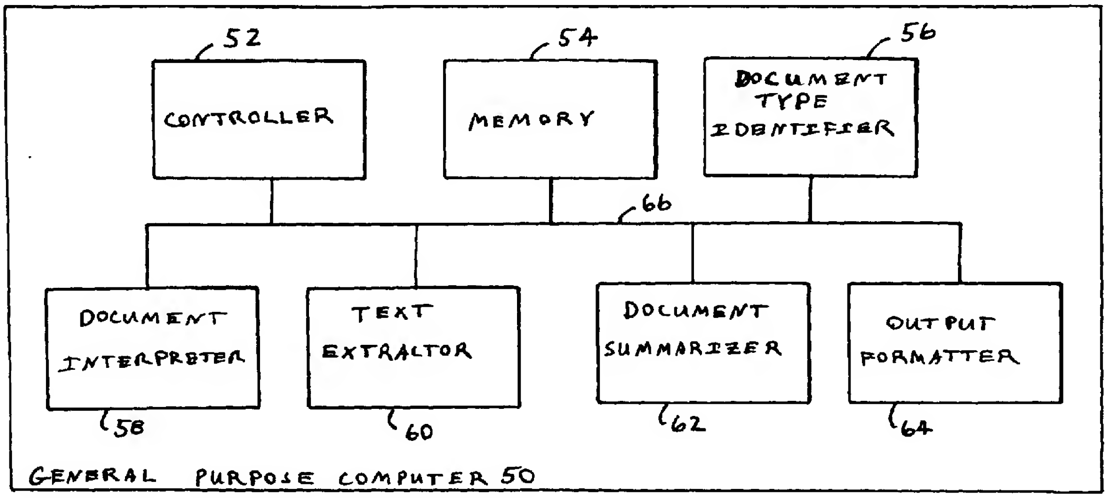


FIG. 3

Printed by Jouve, 75001 PARIS (FR)

Description

[0001] This invention relates to a method and a system for automatic text processing. In particular, this invention relates to a method and system for generating a document summary with document navigation information.

[0002] Automatically generated document summaries serve a valuable function by reducing the time required to review documents. Conventional summaries, while providing a good indication of the total content of the summarized document, do not provide any means for accessing the content of the document. Therefore, if a reader is particularly interested in a certain portion of the summary, the reader is forced to scan the text of the entire document to locate the portions of the document that correspond to points of interest in the summary. There exists a need for assisting a reader in locating portions of a document that correspond to portions of the summary.

[0003] In accordance with one aspect of the present invention, a method for summarizing a document comprises:

extracting text from the document along with corresponding location information;
identifying portions of the extracted text that reflect the content of the document;
generating a presentation file that includes the identified portions and a first set of indicators that indicate the locations of the corresponding extracted text in the document; and
presenting the presentation file.

[0004] Preferably, the method further comprises formatting the layout of a presentation file that includes the identified portions and a first set of indicators that indicate the locations of the corresponding extracted text in the document

[0005] In accordance with a second aspect of the present invention, an automatic document summarizer comprises:

a processor that comprises:
a text extractor that extracts text from a document along with corresponding extracted text location information;
a document summarizer that identifies portions of the extracted text that reflect the content of the document; and
a presentation file formatter that formats the layout of a presentation file that includes the identified portions and a first set of indicators that indicate the locations of the corresponding extracted text in the document; and
a presentation system that presents the presentation file.

[0006] This invention generates a summary page that includes indicators that help the user find the corresponding place in the summarized document from which the summary information was extracted or to which the summary information is most related. This invention enables the user to navigate from the summary to the related locations in the document. This invention also highlights the portions of the document that have been extracted for the summary.

[0007] This invention connects the summary information with the contents of the summarized document by adding location information and/or navigation information to the summary. The method and the system of this invention work with conventional image or text document summarization systems by preserving the location information of the content that is extracted from a document and by generating a summary with indicators that indicate the corresponding locations in the document from which each portion of the summary was extracted. In one embodiment, these indicators indicate page and line numbers and have matching indicators within the summarized document.

[0008] This invention prepares a file from the document for input into a conventional document summarizer, receives the output from the automatic document summarizer and conditions the output to provide indicators to the output summary and document.

[0009] The indicators of this invention enable the user to quickly access the corresponding portions of the document that are related to the summary by directly accessing the document using the information indicated by the indicator or by locating a matching indicator within the outputted document. This task is made easier by highlighting the extracted portions within the outputted document.

[0010] This invention uses an automatic summarizer, linked to a device that can identify the page image and page location in the original document where the summary material is located. The location information is generated by a conventional or modified text or image summarizer. In image summarization, a page number can be found by applying image matching methods. In the case of text summarization, conventional printing software can be modified such that the page numbers assigned by it are made known to the software implementing the present invention.

[0011] The preferred embodiments of this invention will be described in detail, with reference to the following figures, wherein:

Fig. 1 is a block diagram of one embodiment of the document summarizer of this invention;
Figs. 2A-2B show a flowchart outlining the control routine of one embodiment of this invention;
Fig. 3 is a block diagram of one embodiment of a processor of this invention.
Fig. 4 is a sample summary generated by one embodiment of this invention; and

Fig. 5 shows the source document that was analyzed to produce the summary of Fig. 4.

[0012] Fig. 1 shows a block diagram of one embodiment of the document summarizer 10 of this invention. The document summarizer 10 includes a processor 12 communicating with a memory 14 and a hard drive 16. The processor 12 also communicates through an input/output interface 18 to any number of conventional input/output devices such as a pen 20, a mouse 22, a keyboard 24, a display 26, an optical character recognition system, a printer 30, and a scanner 32. The input/output devices 20-32 are operated by a user to control the operation of the document summarizer 10. The processor 12 also communicates through link 34 to an external data source 38. The link 34 includes an interruptible connection 36.

[0013] As shown in Fig. 1, the system 10 is preferably implemented using a programmed general purpose computer. However, the system can be implemented using a special purpose computer, a programmed microprocessor or microcontroller and any necessary peripheral integrated circuit elements, an ASIC or other integrated circuit, a hardwired electronic logic circuit such as a discrete element circuit, a programmable logic device such as a PLD, PLA, FPGA, or PAL, or the like. In general, any device on which a finite state machine capable of implementing the flowcharts shown in Figs. 2A and 2B can be used to implement the system 10.

[0014] Additionally, as shown in Fig. 1, the memory 14 is preferably implemented using static or dynamic RAM. However, the memory 14 can also be implemented using a floppy disk and disk drive, a writeable optical disk and disk drive, a hard drive, flash memory, or the like. Additionally, it should be appreciated that the memory 14 can be a distinct portion of a single memory or physically distinct memories.

[0015] Further, it should be appreciated that the link 34 connecting the microprocessor 12 to the external data source 38 can be a wired or a wireless link to a network (not shown). The network can be a local area network, a wide area network, an intranet, the Internet, or any other distributed processing and storage network. In this case, a document may be downloaded from an external data source 38 for processing by the system 10 according to the method outlined below.

[0016] Figs. 2A and 2B show a flowchart outlining the control routine of one embodiment of the method of this invention. Beginning in step 100, the control routine continues to step S110 where a document is input into the system 10. In step S120, the control routine determines whether the document is summarizable. If the document is not summarizable, the control routine continues to step S130. In step S130, the control routine prints a "generic" cover sheet and continues to step S140. The "generic" cover sheet may include any type of information that may not be related to the input document. In step S140, the document is printed and the control routine continues to step S150 where the control routine stops.

[0017] If in step S120, the control routine determines that the document is summarizable, the control routine jumps to step S160. In step S140, the control routine determines if the document is of a special type that requires or may benefit from special processing. A document of a special type may be identified by a user. An example of a special type includes a group of e-mail messages that each have a summary field. This example will be explained in more detail below.

[0018] If the document is in a special format, the control routine continues to step S170. However, if the control routine determines that the document is not in a special format, the control routine jumps to step S170. In step S150, the control routine extracts the summary information in accordance with the special processing and jumps to step S270.

[0019] In step S180, the control routine identifies the format of the document and continues to step S190. In step S190, the control routine generates a printable file from the input document. This printable file is necessary to generate layout information which will be used in subsequent steps. The control routine then continues to step S200 where the control routine extracts text from the printable file and creates another file that is suitable for processing by a summarizer. Then the control routine continues to step S210 where the control routine creates a layout file having position information on the printable file created in step S190 and corresponding pointers to the extracted text file created in step S200. The control routine then continues to step S220, where the control routine analyzes the extracted text file to determine which portions of text will be included in the summary. The control routine may use a conventional or modified text summarizer. The control routine then continues to step S230 where the control routine generates a printable output file by concatenating the summary text in the printable file created in step S190. Preferably, the summary is placed in the output file before the original document in a cover page. The control routine then continues to step S240. In step S240, the control routine modifies the output file by attaching indicators that indicate the location of the portions of the summary in the original document. The indicators are created by referring to the extracted text file and to the layout file. The control routine then continues to step S250 where the output file is modified to place the indicators in the margin of the cover page. The control routine then continues to step S260.

[0020] In step S260, the control routine modifies the output file by adding additional indicators that are positioned adjacent to the portions of the document that correspond to portions of the summary. The control routine also highlights the indicated portions of the document in step S260. The control routine then continues to step S270. In step S270, the control routine sends the output file to a printer to print the document and then continues to step S280 where the control routine stops.

[0021] If the input document is identified as requiring special processing in step S160 then the control routine continues to step S170. In step S170, the control routine generates an output file that has a summary page with indicators that indicate from where each portion of the summary originated in the input document. The intention of steps S160 and S170 are to make clear that one of ordinary skill in the art would know that steps S180-S260 can be modified in accordance with the input document type and still form a part of the invention.

[0022] One example of an input file that could benefit from special processing is a document that includes several e-mail messages that each have a "subject" field. The control routine recognizes that such a document requires or can benefit from special processing. Alternatively, a user can identify the input document as requiring special processing to the control routine with a user interface (not shown). In this example, the processing would take advantage of the knowledge that there are predefined "subject" fields and will generate a summary from these "subject" fields and thereby obviate the necessity of performing a text summarization process on the document. The summary would include the text from each "subject" field along with indicators that indicate the position in the output file where each portion of the summary may be found.

[0023] A specific example of the operation of this invention will now be described while referring to the flowcharts of Figs. 2A and 2B. The source document in this example is a PostScript file. PostScript is a page-description language from Adobe Systems, Inc. that offers flexible font capability and high-quality graphics. PostScript uses English-like commands to control page layout and to load and scale outline fonts. A page description language describes output to a printer or a display device using instructions from the page-description language to construct text and graphics and to, thereby, create the required page image.

[0024] Processing of the PostScript file starts at step S100 and continues to step S110 where the file is input to the system. The control routine continues to step S120 where the control routine determines that the document is summarizable and then continues to step S160 where the control routine determines that the PostScript file is not one of a predetermined special type.

[0025] The control routine continues to step S180 where the format of the input document is identified as PostScript. The control routine continues to step 190 where the control routine generates a printable output file. In this example, because the PostScript file is already printable, the control routine merely makes a copy of the input document as the printable output file. Therefore, the printable output file is also a PostScript file. The control routine then continues to step S200 where the text is extracted from the printable file. The control routine continues to step S200 where the control routine extracts the text using a modified PostScript interpreter and a text extraction program. In step S200, the PostScript interpreter analyzes the PostScript file and generates an ASCII file that includes position information. The invention then utilizes a text extraction program. The text extraction program maps font differences, analyzes the stream of characters, and picks out words, lines, and paragraphs. The text extraction program also records the location information in accordance with the identified, extracted text. The control routine then continues to step S210. In step S210 the control routine generates a layout file and continues to step S220.

[0026] In step S220, the system analyzes the ASCII file using text summarizer to determine what portions of the extracted text are to be included in the summary. The control routine then continues to step S230 where a PostScript generator creates a printable output file. In step S230, the output file is created using the summarized text from step S220 and the position information corresponding to the summarized text from the layout file created in step S210. The control routine then continues to step S240. In step S240, the control routine modifies the output PostScript file with instructions that add indicators to the summary. The control routine then continues to step S250. In step S250, the control routine appends the output PostScript file with instructions that add indicators to the margin of the cover page. The control routine then continues to step S260.

[0027] In step S260, the control routine modifies the output PostScript file to highlight those portions of the document that correspond to the portions of the summary. The control routine then continues to step S270 where the control routine sends the output PostScript file to a printer and then continues to step S280 where the control routine stops.

[0028] It should be appreciated that while the invention has been described above in connection with PostScript and ASCII text files that this invention may be used in conjunction with any type of document format including, for example, any type of graphic, audio, and video document formats.

[0029] It should also be appreciated that steps S210, S220 and S240 are all optional and that the indicators may take whatever form is appropriate for the desired style of output. The indicators only need to provide a reference to the document content by providing position information or as an indicator to be matched with another indicator within the document.

[0030] Additionally, while the embodiment of the control routine described above and in Figs. 2A and 2B generates a single output file, one of ordinary skill understands that the control routine may generate any number of output files such as, for example, a first output file having only the cover page with a summary and a second output file having only the document.

[0031] Typical document summarizers present the sentences in the summary in the same order that the sentences are encountered in the summarized document. However, these sentences may not be in the same order in the summary as they appear in the related document. In that case, the indicators may be presented in the summary in the order that the summary presents the sentences or in the order in which the sentences appear in the related document. If the indicators are to appear on the summary page in the order that the sentences appear in the related document, then leaders may be provided from the indicators to each of the corresponding portions of the summary.

[0032] The output of the summarized document may not facilitate easy positioning of indicators to readily indicate which portions of the summarized document correspond to the indicator. Therefore, leaders may also be added from the indicators in the summarized document to the portions of the summarized document that have been extracted for the summary.

[0033] The indicators may be placed on the summary page in a vertical position on the margin that corresponds to the vertical position of the corresponding portions of the summarized document. Such an arrangement makes it easier for a user to find an indicator within the summarized document when flipping through a document. Indicators in the summarized document may be eliminated very quickly when the vertical position does not match the vertical position of the indicator in the summary sheet. On the other hand, an indicator that has a vertical position that matches the vertical position of the corresponding indicator in the summary sheet can quickly indicate the correspondence between the two indicators to the user of the document.

[0034] Additionally, this invention also provides indicators to a list of keywords. Conventional document cataloging systems generate a list of keywords that may be quickly searched to locate a document related to the keyword. This invention is also useful with a keyword list generating system to provide indicators adjacent the keywords to indicate the location of the keywords within the summarized document.

[0035] The indicators may also take the form of a background color or font color on a portion of the document summary. In this case, a matching indicator in the summarized document is highlighted using a background or font color that matches the corresponding portion of the summary. The indicators may also take the form of character attributes such as Bold or Italics. In this case, for example, a portion of the summary may be bold while another portion is italicized. The correspondence to the original document would be indicated because the source text would have a corresponding character attribute. It is to be understood that the indicators may take any form that allows one to distinguish between indicators and also indicates the correspondence of each portion of the summary to the source in the underlying original document.

[0036] Fig. 3 shows a block diagram of one embodiment of the processor 12 of this invention. The processor 12 is preferably implemented using a general purpose computer 50. The general purpose computer 50 preferably includes a controller 52, a memory 54, a document type identifier 56, a document interpreter 58, a text extractor 60, a document summarizer 62, an output formatter 64, and a bus 66. The elements of the general purpose computer 50 are interconnected by the bus 66. The document type identifier 56, the document interpreter 58, the text extractor 60, the document summarizer 62, and the output formatter 64 are used to implement the flowchart of Figs. 2A and 2B. It should be appreciated that the document type identifier 56, the document interpreter 58, the text extractor 60, the document summarizer 62, and the output formatter 64 are preferably implemented as software routines running on the controller 52 and stored in the memory 54. It should also be appreciated that many other implementations of these elements will be apparent to those skilled in the art.

[0037] It should be understood that the term "document" is intended to include text, audio, video, or any other information storing file in any combination of information storing files. Further, it should be understood that the term "text" is intended to include text, digital ink, audio, audio bars, video, or any other content of a document, including a document's structure. It should also be appreciated that the term "display" is intended to include any type of presentation device appropriate for the type of text in the document. It should be further understood that the term "portion" is intended to include any divisible structure of a document such as, for example, words, sentences, phrases, paragraphs, sections, pages, and any other distinguishable portion of a document. Structural information designates divisible portions of documents.

[0038] Fig. 4 shows a sample output of one embodiment of this invention. Fig. 5 shows the source document 70 which has been summarized.

[0039] A summary sheet 72 includes a header 74 and a summary 76. The summary sheet 72 also includes the title of the summarized document 78 adjacent to an extracted portion of the document 80 that generally reflects the overall content of the document. The summary 76 includes extracted portions 82 which reflect the content of portions of the summarized document 70 and a list of keywords 88. Indicators 84 have been placed adjacent to the corresponding extracted portions 82. These indicators 84 have corresponding tabs 86 positioned on the margin of the summary sheet 72. The tabs 86 indicate the vertical position of the extracted portion in the summarized document 70 that correspond to the indicators 84. In this example, the summarized document has not been output with indicators or tags that correspond to the indicators 84 and tags 86 of the summary sheet. It should be appreciated that other embodiments of this invention may provide indicators and tabs in the output of the summarized document 70.

[0040] While the above explanation and description has generally referred to position and location information, it is to be understood that the indicators of this invention include valuable navigation information as a result of the fact that the indicators include position information. Therefore, the indicators include navigation information which is helpful to a user to navigate a document to locate the position in a document from which a particular portion of summary originated.

Claims

1. A method for summarizing a document, the method comprising:

   extracting text from the document along with corresponding location information;
   identifying portions of the extracted text that reflect the content of the document;
   generating a presentation file that includes the identified portions and a first set of indicators that indicate the locations of the corresponding extracted text in the document; and
   presenting the presentation file.

2. The method of claim 1, where the location information comprises structural information.

3. The method of claim 1 or claim 2, where the location information comprises layout information.

4. The method of any of claims 1 to 3, further comprising the step of identifying the document as being of a predetermined type, wherein the steps of extracting and identifying correspond to the predetermined type.

5. The method of any of claims 1 to 4, further comprising:

   generating a second set of indicators that correspond to the first set of indicators; and
   placing the second set of indicators in the layout of the document adjacent to the corresponding extracted text, wherein the first set of indicators comprise presenting each corresponding portion of summary in a presentation attribute that corresponds to the second set of indicators that comprise the presentation of the corresponding extracted text of the document in the same corresponding presentation attribute.

6. An automatic document summarizer comprising:

   a processor that comprises:
   a text extractor that extracts text from a document along with corresponding extracted text location information;
   a document summarizer that identifies portions of the extracted text that reflect the content of the document; and
   a presentation file formatter that formats the layout of a presentation file that includes the identified portions and a first set of indicators that indicate the locations of the corresponding extracted text in the document; and
   a presentation system that presents the presentation file.

7. The summarizer of claim 6, where the location information comprises structural information.

8. The summarizer of claim 6 or claim 7, where the location information comprises layout information.

9. The summarizer of any of claims 6 to 8, wherein the processor further comprises a document type identifier that identifies the document as being of a predetermined type, wherein the text extractor extracts the text using a method corresponding to the identified predetermined type and the document summarizer identifies portions using a method corresponding to the identified predetermined type.

10. The summarizer of any of claims 6 to 9, where the presentation file formatter generates a second set of indicators that correspond to the first set of indicators and places the second set of indicators in the layout of the document adjacent to the corresponding extracted text, wherein the first set of indicators comprise presenting each corresponding portion of summary in a presentation attribute that corresponds to the second set of indicators that comprise the presentation of the corresponding extracted text of the document in the same corresponding presentation attribute.

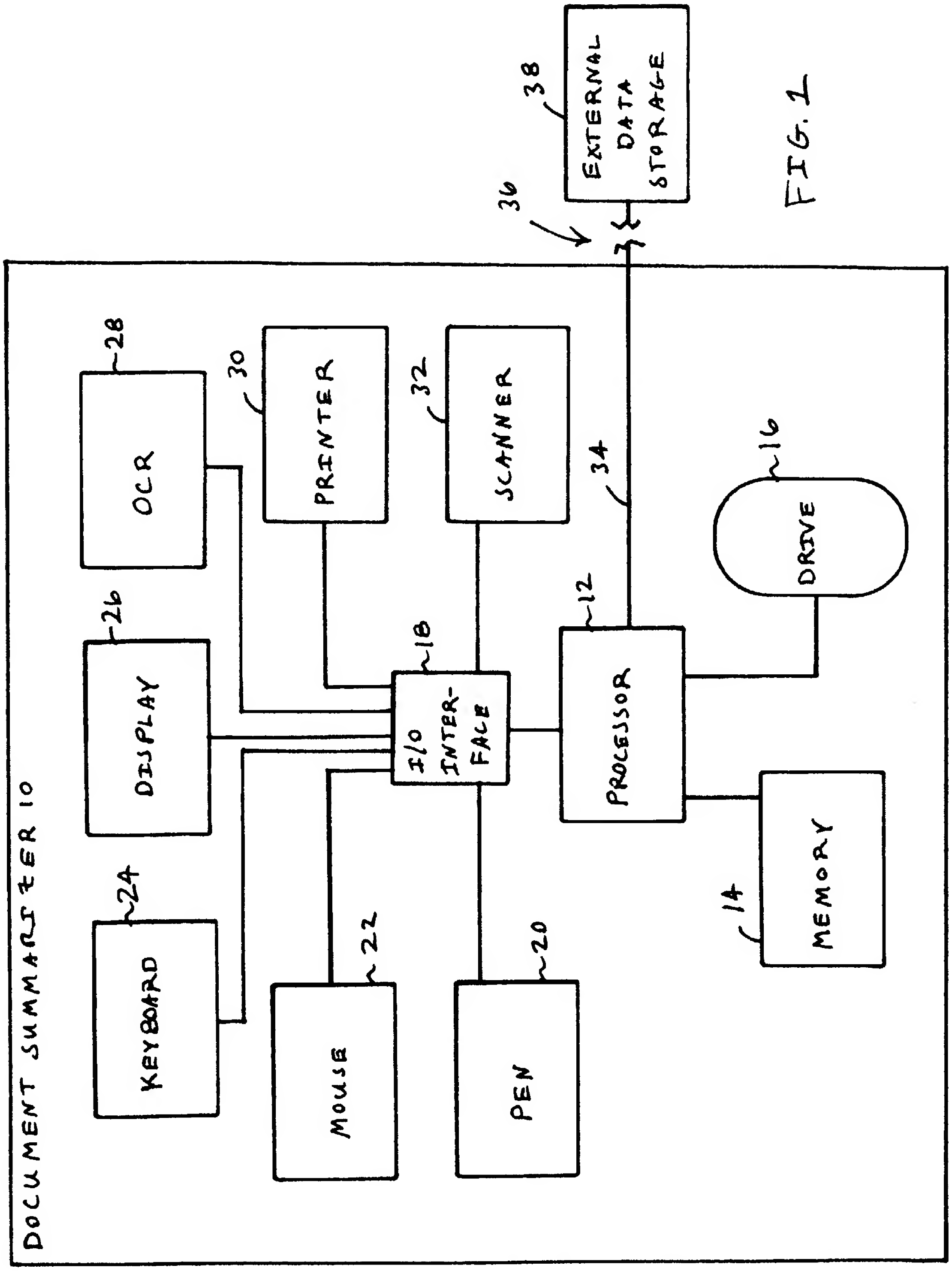

DOCUMENT SUMMARIZER 10
KEYBOARD
24
MOUSE
22
PEN
20
DISPLAY
26
OCR
28
PRINTER
30
SCANNER
32
I/O INTER-FACE
18
PROCESSOR
12
MEMORY
14
DRIVE
16
34
36
EXTERNAL DATA STORAGE
38

FIG. 1

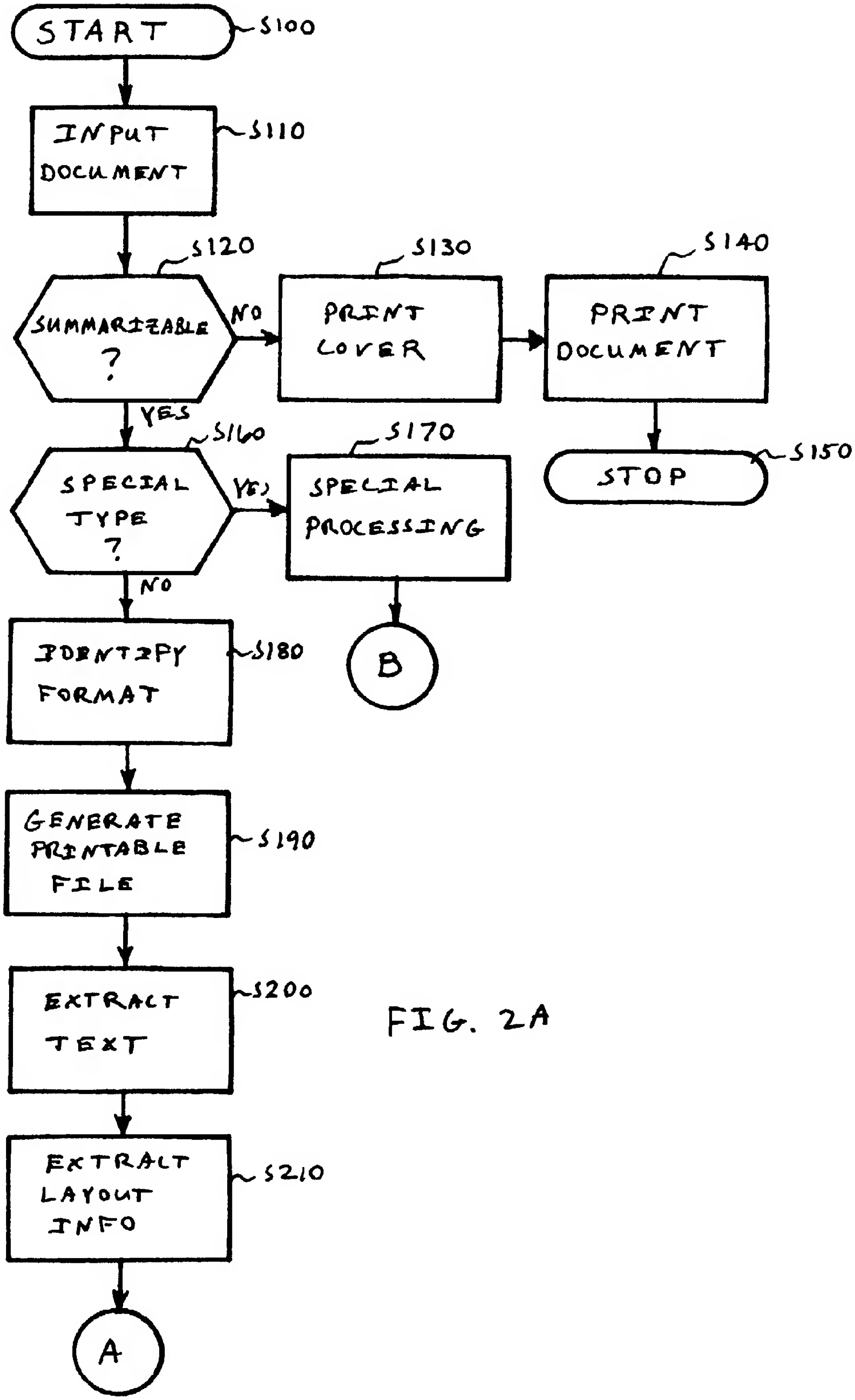

START
S100
INPUT DOCUMENT
S110
S120
SUMMARIZABLE ?
NO
S130
PRINT COVER
S140
PRINT DOCUMENT
S150
STOP
YES
S160
SPECIAL TYPE ?
YES
S170
SPECIAL PROCESSING
B
NO
IDENTIFY FORMAT
S180
GENERATE PRINTABLE FILE
S190
EXTRACT TEXT
S200
EXTRACT LAYOUT INFO
S210
A

FIG. 2A

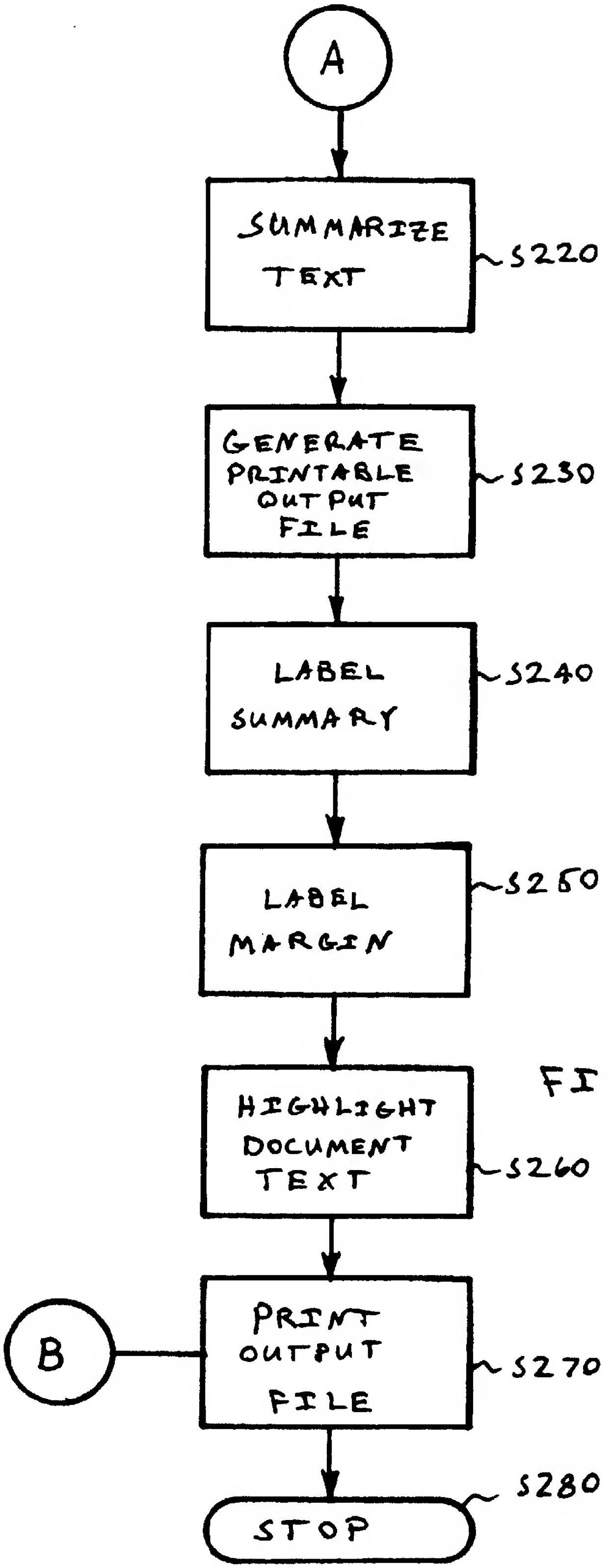

A
SUMMARIZE TEXT
S220
GENERATE PRINTABLE OUTPUT FILE
S230
LABEL SUMMARY
S240
LABEL MARGIN
S250
HIGHLIGHT DOCUMENT TEXT
S260
B
PRINT OUTPUT FILE
S270
S280
STOP

FIG. 2B

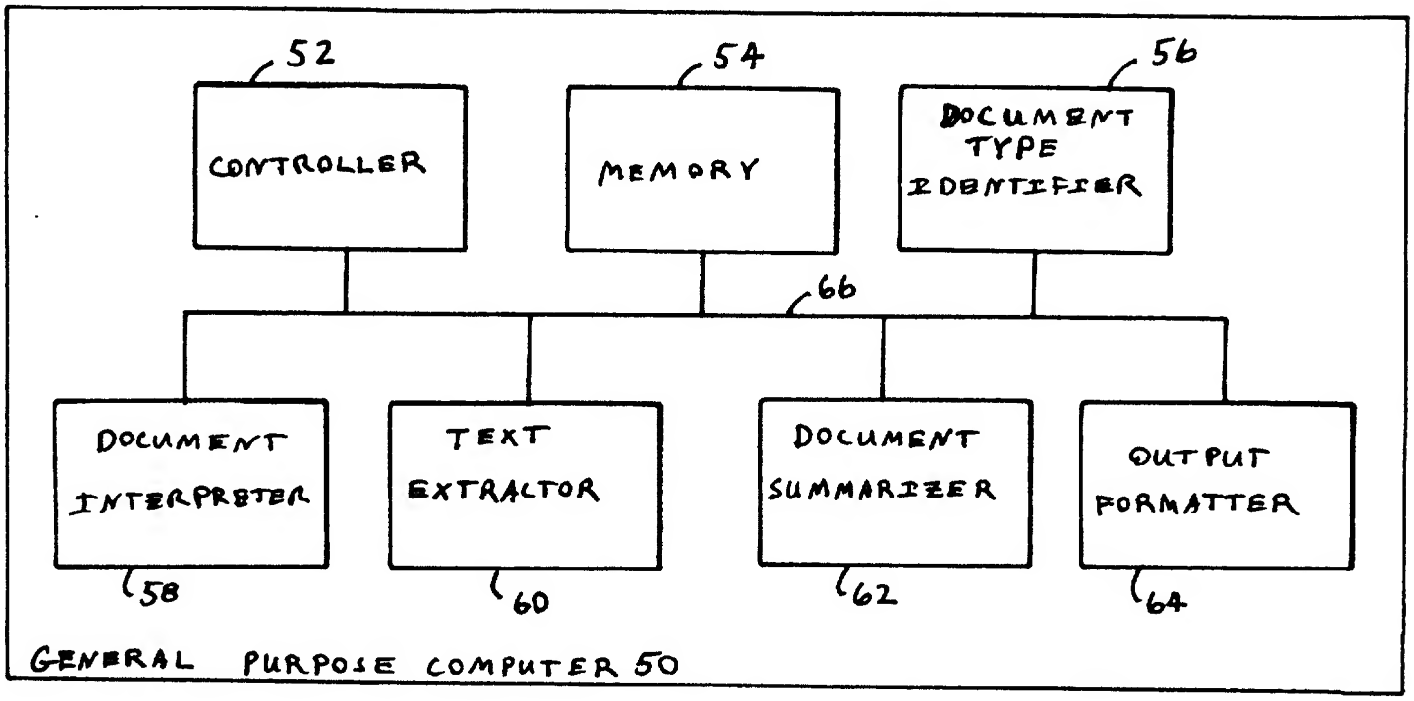
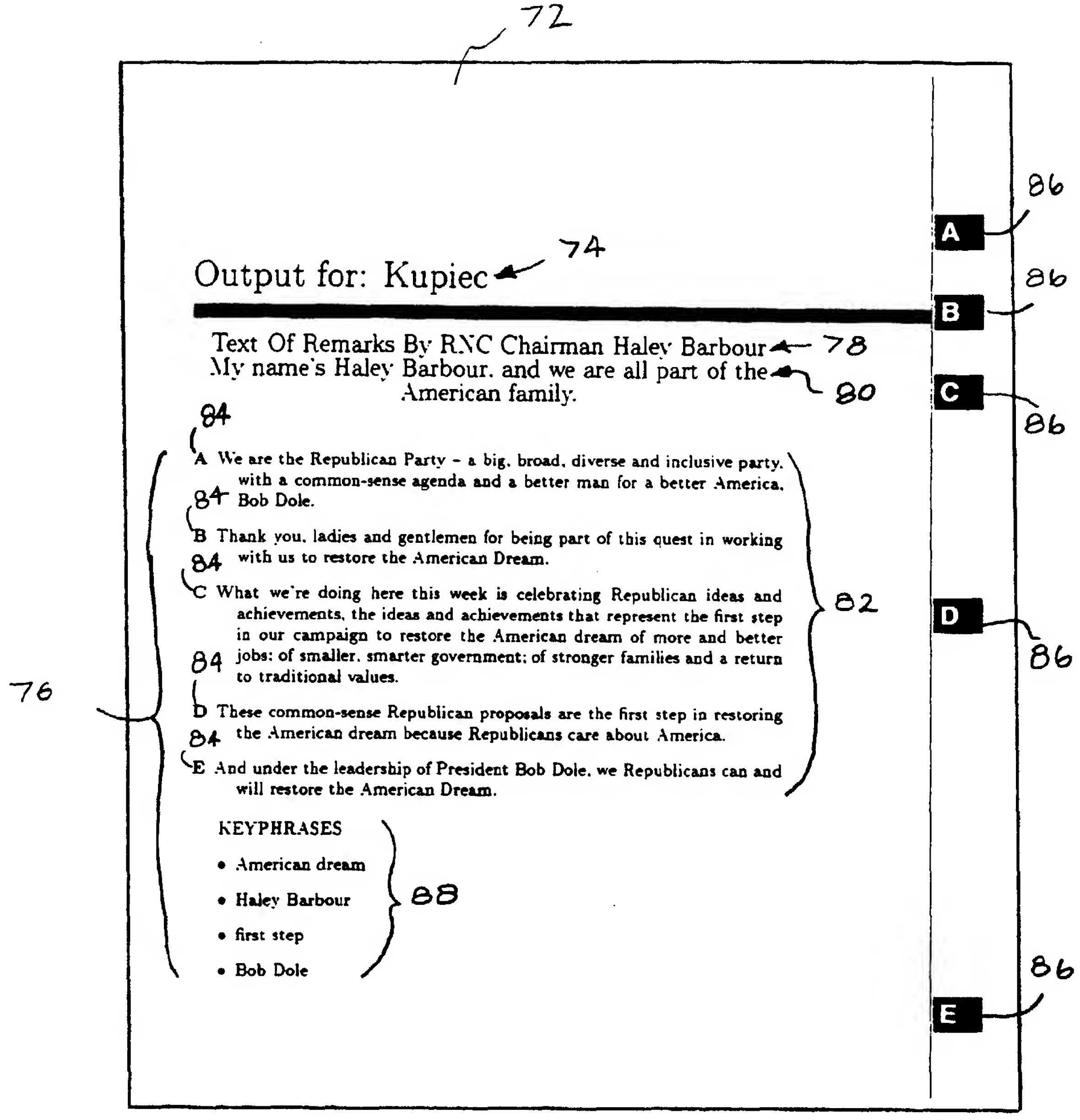

52
CONTROLLER
54
MEMORY
56
DOCUMENT TYPE IDENTIFIER
66
DOCUMENT INTERPRETER
58
TEXT EXTRACTOR
60
DOCUMENT SUMMARIZER
62
OUTPUT FORMATTER
64
GENERAL PURPOSE COMPUTER 50

FIG. 3



Output for: Kupiec

Text Of Remarks By RNC Chairman Haley Barbour
My name's Haley Barbour, and we are all part of the American family.

A We are the Republican Party - a big, broad, diverse and inclusive party, with a common-sense agenda and a better man for a better America, Bob Dole.

B Thank you, ladies and gentlemen for being part of this quest in working with us to restore the American Dream.

C What we're doing here this week is celebrating Republican ideas and achievements, the ideas and achievements that represent the first step in our campaign to restore the American dream of more and better jobs; of smaller, smarter government; of stronger families and a return to traditional values.

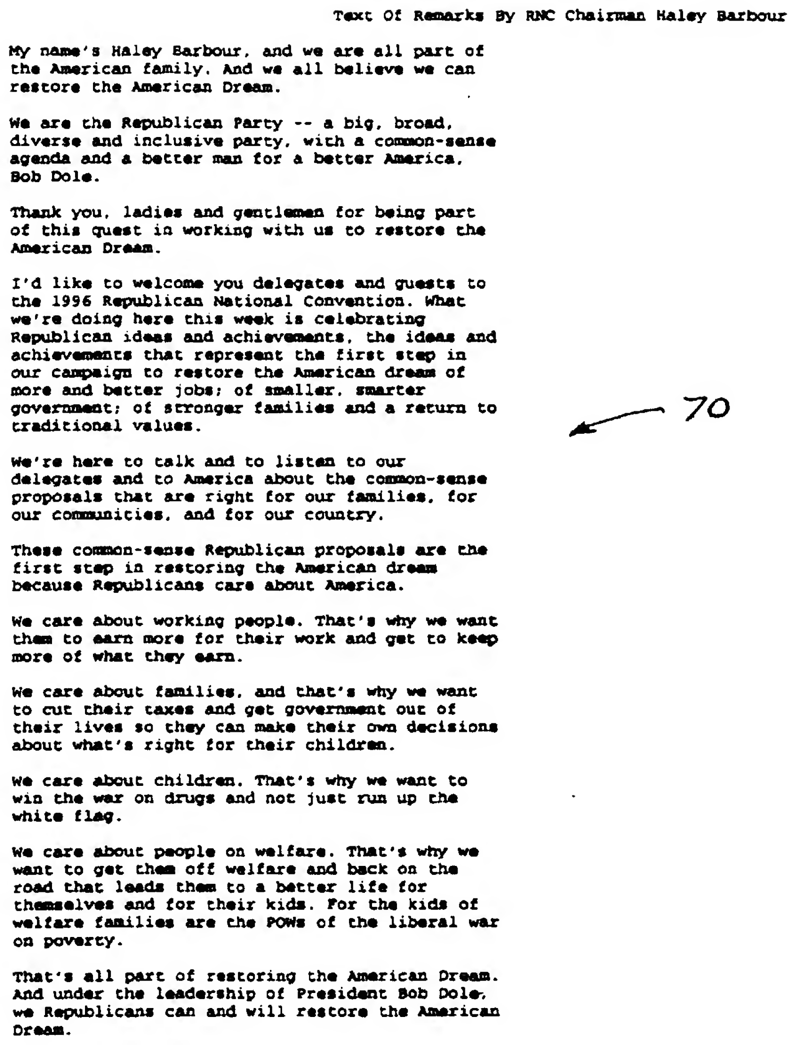
D These common-sense Republican proposals are the first step in restoring the American dream because Republicans care about America.

E And under the leadership of President Bob Dole, we Republicans can and will restore the American Dream.

KEYPHRASES

- American dream
- Haley Barbour
- first step
- Bob Dole

FIG. 4



Text Of Remarks By RNC Chairman Haley Barbour

My name's Haley Barbour, and we are all part of the American family. And we all believe we can restore the American Dream.

We are the Republican Party -- a big, broad, diverse and inclusive party, with a common-sense agenda and a better man for a better America, Bob Dole.

Thank you, ladies and gentlemen for being part of this quest in working with us to restore the American Dream.

I'd like to welcome you delegates and guests to the 1996 Republican National Convention. What we're doing here this week is celebrating Republican ideas and achievements, the ideas and achievements that represent the first step in our campaign to restore the American dream of more and better jobs; of smaller, smarter government; of stronger families and a return to traditional values.

We're here to talk and to listen to our delegates and to America about the common-sense proposals that are right for our families, for our communities, and for our country.

These common-sense Republican proposals are the first step in restoring the American dream because Republicans care about America.

We care about working people. That's why we want them to earn more for their work and get to keep more of what they earn.

We care about families, and that's why we want to cut their taxes and get government out of their lives so they can make their own decisions about what's right for their children.

We care about children. That's why we want to win the war on drugs and not just run up the white flag.

We care about people on welfare. That's why we want to get them off welfare and back on the road that leads them to a better life for themselves and for their kids. For the kids of welfare families are the POWs of the liberal war on poverty.

That's all part of restoring the American Dream. And under the leadership of President Bob Dole, we Republicans can and will restore the American Dream.

FIG. 5